CAPSTONE PRESENTATION

SENTIMENT ANALYSIS
“Neural Network - Deep Learning”

Collaborator: Luka Anicin
Data Scientist: Ruslan S.

OBJECTIVES

Problem: predict the company’s image based on users’ reviews

Context: the "face" of the company plays a profound role in its success. It's reflected in its reputation, trustworthiness, accountability, and other important aspects that drives the business’ success

METRICS

- Model accuracy: 65-89%
- Delivery: October 2021

CONSTRAINTS

- Data
- Time
- Limited stuff

STAKEHOLDERS

- Business with twitter account

PROBLEM IDENTIFICATION

- Approach:
  Supervised Classification
  (2 classes)

- Learning algorithm:
  CNN Sequential
  (5 hidden layers)

DATA DESCRIPTION

Column names: “target” and “predictor”
Dimension: (1,583,691 by 2)
Date/time range: 04/06/2009-06/16/2009

```
Unique values:  ['Negative' 'Positive']
     target                                                predictor
0  Negative  @switchfoot http://twitpic.com/2y1zl - Awww, t...
1  Negative  is upset that he can't update his Facebook by ...
2  Negative  @Kenichan I dived many times for the ball. Man...
3  Negative    my whole body feels itchy and like its on fire
4  Negative  @nationwideclass no, it's not behaving at all....
```

WRANGLING STEPS

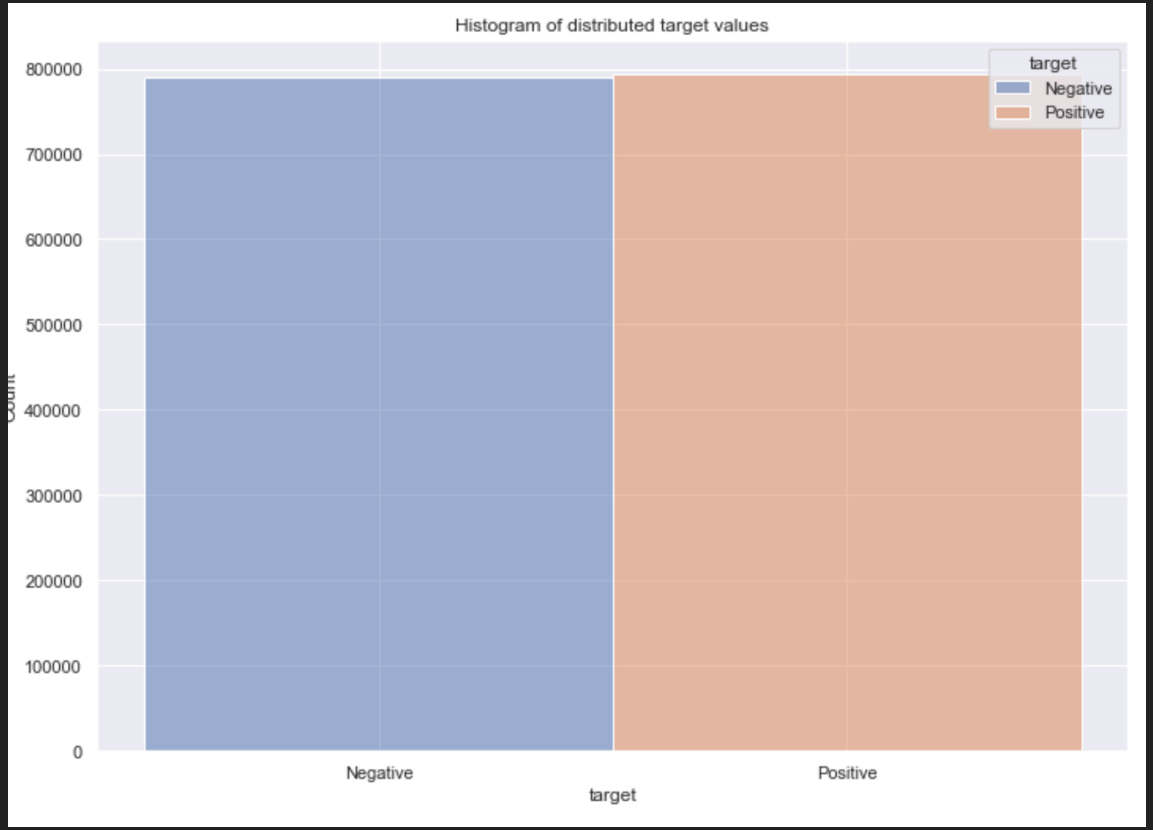
- Dropping irrelevant columns
  - Id, date, query, user_name
- Converting target class into binary with type “int”
  - Values 0 (negative) and 1 (positive)
- Dealing with duplicates
- Removing noise in text (using regular expression)
- Dealing with stop words (using internal library)
- Applying stemming
- Splitting into training and testing datasets
- Tokenization

KEY TAKEAWAYS

- Working with text (in our case with tweets) gets complicated because of dealing with:
  - Grammar mistakes
  - The use of short (simplified) words
  - Combination of different types of characters
  - Emojis
  - And so on...

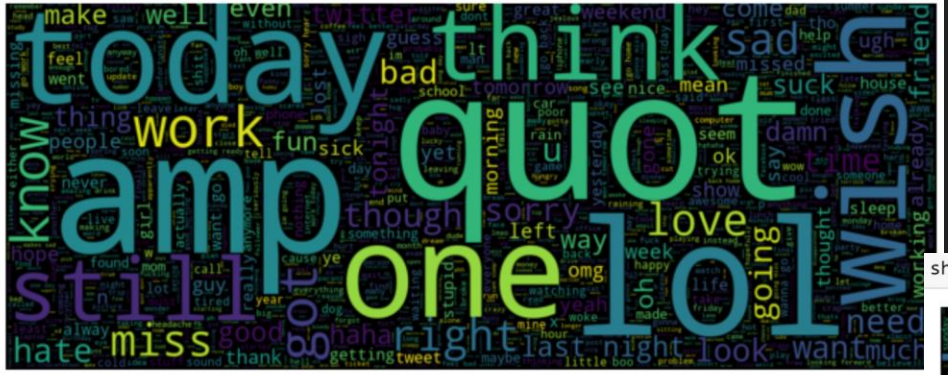
EDA (EXPLORATORY DATA ANALYSIS)

Ratio of classes

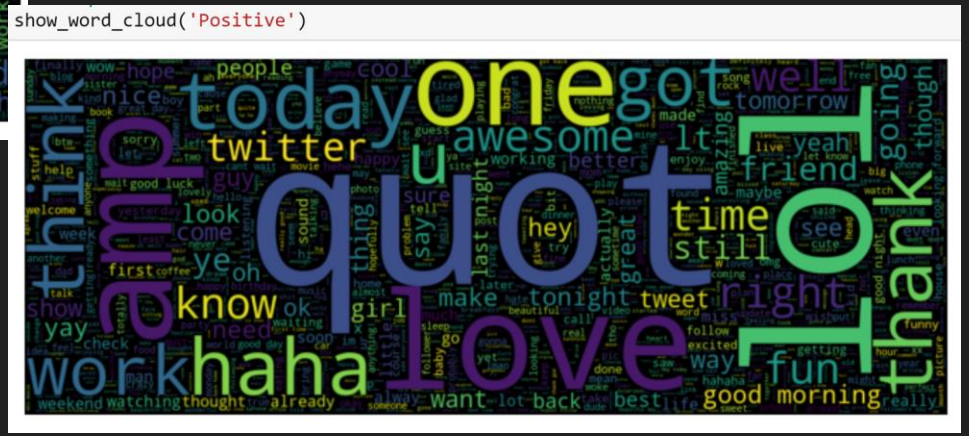


Most common negative and positive words

```
show_word_cloud('Negative')
```




```
show_word_cloud('Positive')
```



MODELING

LEARNING ALGORITHM

- Model type:
  - Sequential
- Architecture:
  - CNN (convolutional neural network 1 dimension)
- Layers:
  - 5 hidden layers: Embedding, SpetialDropout1D, LSTM, and 2 Dense layers

ALGORITHM SUMMARY

Layer (type)	Output Shape	Param #
embedding_3 (Embedding)	(None, 30, 100)	29041200
spatial_dropout1d_2 (Spatial	(None, 30, 100)	0
lstm_2 (LSTM)	(None, 100)	80400
dense_7 (Dense)	(None, 15)	1515
dense_8 (Dense)	(None, 1)	16

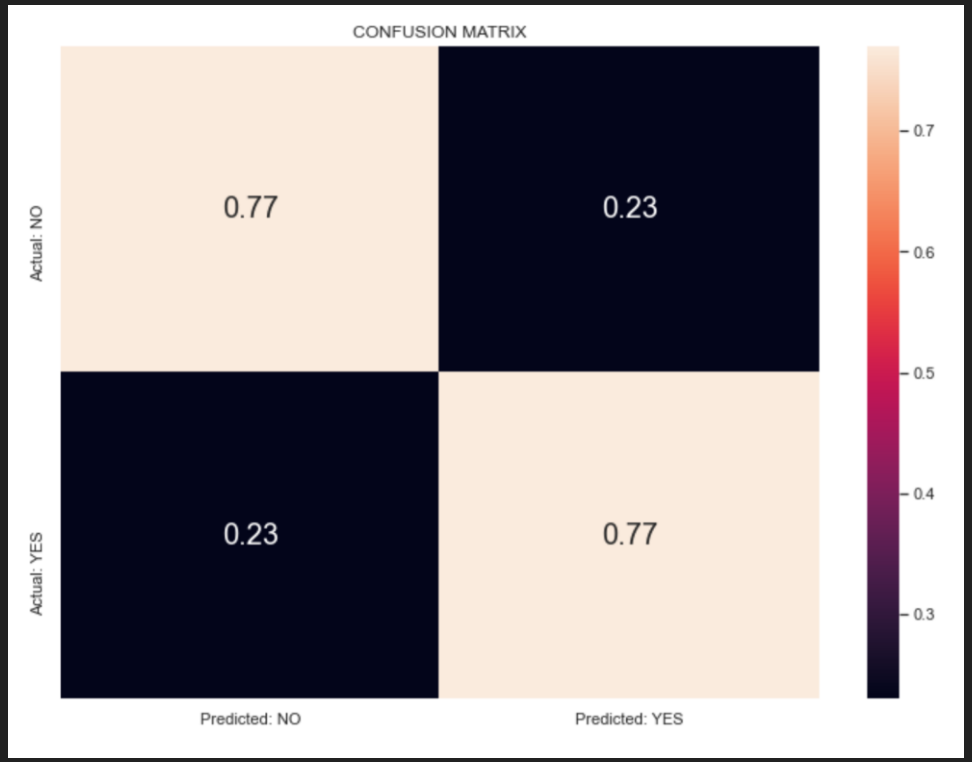
Total params: 29,123,131
Trainable params: 29,123,131
Non-trainable params: 0

FACTORS TOWARD CHOSEN MODEL

- Size of data set
- Accuracy

PERFORMANCE RESULTS

CONFUSION MATRIX



PREDICTION SCORE

	precision	recall	f1-score	support
Class 0 - Negatives	0.77	0.77	0.77	157487
Class 1 - Positives	0.77	0.77	0.77	159252
accuracy			0.77	316739
macro avg	0.77	0.77	0.77	316739
weighted avg	0.77	0.77	0.77	316739

RECOMMENDATION

Model_name	Precision	Recall	F1-score	Accuracy
Model_1	0.75	0.75	0.75	0.75
Model_3	0.75	0.75	0.75	0.75
Model_4	0.77	0.77	0.77	0.77

- Slow to train
- At risk of overfitting

FUTURE WORK

- Increase size
- Use most recent data
- Play with hyper-tuning
- Reconstruct architecture of the model
- Use more libraries to deal with text (to correct grammar mistakes or convert emojis into words and so on)

End of slide show